Probabilistic Models for Spatially Aggregated Data(Abstract_要旨)

AUTHOR(S):

Tanaka, Yusuke

様式VI

博士学位論文調査報告書

論 文 題 目　Probabilistic Models for Spatially Aggregated Data
（空間集約データのための確率モデル）

申請者氏名　田中佑典

最 終 学 歴　平成２５年　３月
京都大学大学院情報学研究科システム科学専攻修士課程　修了
令和　２年　３月
京都大学大学院情報学研究科システム科学専攻博士後期課程
研究指導認定見込

学 識 確 認　令和　　年　　月　　日（論文博士のみ）

論文調査委員　京都大学大学院情報学研究科
（調査委員長）　教　授　田中利幸

論文調査委員　京都大学大学院情報学研究科
教　授　石井　信

論文調査委員　京都大学大学院情報学研究科
教　授　下平英寿

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	田中佑典
論文題目	Probabilistic Models for Spatially Aggregated Data （空間集約データのための確率モデル）		

（論文内容の要旨）

人々の行動や社会活動などに関して，大規模なデータの収集と活用とが行われるようになっているが，個人情報の保護やデータ収集にかかる費用などの様々な観点から，これらのデータは何らかの形で集約される場合が少なくない．このようなデータを，本研究では集約データ（aggregated data）と呼んでいる．集約データは，集約の過程でしばしばデータの活用に際して重要な情報が失われてしまうため，素朴な手法ではデータの活用に支障が生じうる．本学位論文は，空間的に集約されたデータ（空間集約データ）の活用に際して生じうるこのような問題点を解消するために，データ生成過程に関する確率モデルを構築し，それにもとづいたデータ解析手法を研究した成果を取りまとめたものである．

本学位論文では，具体的に2つのタスクを研究対象として取り上げている．ひとつは，空間的に粗い粒度で集約されたデータを高解像度化するタスクであり，もうひとつは，観測領域ごとに得られた人の流入数，流出数のデータから人流を推定するタスクである．対応して，本学位論文は三部構成となっており，第一部ではデータの高解像度化について，第二部では人流推定について，それぞれ議論がなされている．第三部は論文全体の結論に充てられている．

本学位論文の第一章は序論であり，研究の動機および目的を述べるとともに，本学位論文で取り上げる2つのタスクおよび本学位論文の主要な成果について概略を述べている．

第一部は第二章から第五章までからなり，粗い粒度で集約されたデータの高解像度化に関する研究の成果が述べられている．第二章では，まず導入がなされる．同一地域に対して様々な種類のデータが様々な粒度で得られていることを前提とし，それらを活用して高解像度化を行うという問題を扱うことが述べられる．また，既存研究が回帰アプローチと多変量アプローチとに大別されることを指摘するとともに，問題の定式化，ならびに第三章以降で述べられる本研究の概要が提示されている．

第三章では，以降の議論の基礎となる，単一種類のデータに対するモデルが導入されている．ここで導入されるモデルSAGP-Sは，ガウス過程と空間的集約に対応する観測過程とを組み合わせたモデルである．SAGP-Sにおいては，空間集約データが得られた際の事後確率モデルがやはりガウス過程になることが示され，事後ガウス過程の平均関数，共分散関数を与える式が導出されている．また，ハイパーパラメータ推定に際して重要な周辺尤度の式も示されている．

第四章では，本学位論文でのひとつめの提案手法である2段階SAGPモデルが議論されている．2段階SAGPモデルは回帰アプローチの一例とみなせる．その第一段階では，多種の補助データをそれぞれSAGP-Sでモデル化する．第二段階では，補助データからターゲットとするデータへの回帰係数を推定するが，その際に，ターゲットデータの

空間的集約により粗い粒度のデータが得られるという制約条件を考慮する．ニューヨーク市およびシカゴ市が一般に公開している行政データに提案手法を適用し，提案手法の有効性を示している．
第五章では，本学位論文でのふたつめの提案手法であるSAGP-Mが議論されている．SAGP-Mは多変量アプローチの一例とみなすことができる．ターゲットデータおよび補助データを，いくつかの独立な潜在ガウス過程の線形結合によってモデル化し，SAGP-Sと同様に空間的集約に対応する観測過程とを組み合わせる．空間集約データが得られた際の事後確率モデルがSAGP-Sと同様にやはりガウス過程になることが示され，事後ガウス過程の平均関数，共分散関数を与える式が導出されている．また，ハイパーパラメータや線形結合の係数の推定に際して重要な周辺尤度の式も示されている．さらに，複数地域を同時に考慮することにより，入手可能なデータの種類が少ない地域におけるデータの高解像度化を，多種のデータが入手できる地域のデータと組み合わせて行う手法が提案されている．ニューヨーク市およびシカゴ市が一般に公開している行政データに提案手法を適用し，提案手法の有効性を示している．2段階SAGPモデルとSAGP-Mとを比較すると，高解像度化の精度では後者が優れているのに対し，前者は計算負荷が相対的に低く，実際の状況における精度と計算負荷とのトレードオフに応じていずれかの手法を選択できる，ということが述べられている．
第二部は第六章からなり，観測領域ごとに得られた人の流入数，流出数のデータから人流を推定するタスクに関する研究の成果が述べられている．都市，展示会場，大規模商業施設などの対象領域に複数の観測領域を設定し，各観測領域において時刻ごとの人の流入数，流出数のデータは得られるが，個人を追跡する情報は得られないという状況を考え，観測領域間の人流を推定するという問題が検討されている．既存研究は，流入数，流出数に関する保存則が厳密に成立するという仮定にもとづいているが，本学位論文では観測領域が対象領域を被覆していない状況を想定し，推定されるべき領域間の人流を潜在変数として，保存則を確率的にモデル化するとともに観測領域間の移動にかかる時間の分布も明示的にモデル化する手法が提案されている．また，人流の推定とモデルパラメータの推定を，近似EMアルゴリズムによって同時推定する手法も併せて提案されている．展示会場における参加者のデータおよびニューヨーク市におけるタクシー・バイクシェアの利用データに本研究での提案手法を適用し，既存手法と比較して提案手法が人流をより高い精度で推定できたこと，および移動時間分布の推定も適切になされたことを示す実験結果が述べられている．
第三部は第七章からなり，学位論文全体の結論が述べられている．第一部，第二部で得られた研究成果がまとめられるとともに，研究の今後の展開に関する展望が述べられている．

（続紙 ２ ）

（論文審査の結果の要旨）

ライフログや行政データをはじめとして人々の行動や社会活動に関するデータの蓄積が進み，それらの一層の活用が望まれる．その一方で，個人情報保護やデータ収集に要するコストなどの観点から，これらのデータは何らかの形で集約されることが少なくなく，このことがデータの高度な分析を妨げる大きな要因ともなっている．本学位論文は，空間的に集約されたデータ（空間集約データ）に関する2つのタスクを取り上げ，それぞれに対して確率モデルを構築し，構築された確率モデルに基づいたデータ解析手法を研究した成果を取りまとめたものである．本研究で得られた主な成果は以下の通りである．

1. 空間データの統計的分析によく用いられるガウス過程においては，観測データは空間内の点において得られたものと仮定され，そのままでは空間集約データの分析に使うことができない．本論文では，ガウス過程にもとづく枠組みで空間集約データの分析を可能とするために，ガウス過程と空間的集約に対応する観測過程とを組み合わせたモデルSAGP-Sの定式化を示している．空間的集約に対応する観測過程をリーマン和の極限として定式化し，集約データが得られた際の周辺尤度および事後ガウス過程を明示的に導出している．
2. 多種の空間集約データを活用して空間的に粗い粒度の集約データを高解像度化するタスクに関して，本論文では，上記のSAGP-Sを基礎として回帰アプローチにもとづく手法である2段階SAGPモデル，ならびに多変量アプローチにもとづく手法SAGP-Mを提案している．前者の手法は，高解像度化の対象とする標的データ以外の補助データをまず個別にSAGP-Sで高解像度化し，次に高解像度化された補助データから標的データへの回帰係数を推定することで，補助データと標的データとの相関を利用して高解像度化を行う．後者の手法は，多種のデータをいくつかの独立な潜在ガウス過程の線形結合とみなし，SAGP-Sと同様に空間的集約に対応する観測過程と組み合わせることで高解像度化を行う．ニューヨーク市およびシカゴ市が一般公開している行政データにこれらの手法を適用し，既存手法と比較してより高精度の高解像度化が行えることを実験的に示している．また，2つの提案手法を比較すると，計算量の観点では前者の手法が，精度の観点では後者の手法がそれぞれ優れており，両提案手法は計算量と精度とのトレードオフに応じた使い分けが可能である．
3. 都市，展示会場，大規模商業施設などの対象領域に複数の観測領域を設定し，観測領域ごとの人の流出，流入が流出数，流入数として集約されたデータから観測領域相互間の人流を推定するタスクに関して，本論文では流入数，流出数に関する保存則が厳密には成立しない状況を想定し，さらに観測領域間の移動時間分布を考慮することで既存研究を拡張した定式化を示し，近似EMアルゴリズムにもとづき人流とモデルパラメータとの同時推定を行う手法を提案している．展示会場における参加者のデータおよびニューヨーク市のタクシー・バイクシェアの利用データに適用することで，提案手法の有効性を確認している．

以上に述べたように，本論文は確率モデルにもとづく空間集約データの分析に関して，既存研究をふまえての体系的な定式化，およびそれにもとづいた有用な手法の提案を行ったものであり，データ解析技術をはじめとするデータ駆動型の情報学的手法の発展に資するものである．また，本論文は論理的に明確に記述されており，審査により，申請者は関連事項について高い学識を有するものと判断された．よって、本論文は博士（情報学）の学位論文として価値あるものと認める．令和２年２月１７日，論文内容とそれに関連した事項について試問を行った結果，合格と認めた．併せて，本論文のインターネットでの全文公表ついても支障がないことを確認した．